基于主题模型的 BBS 话题演化趋势分析①

曹丽娜，唐锡晋
（中国科学院数学与系统科学研究院，北京 100190）

摘要： 互联网引发的舆情问题愈发突出，网络舆情研究已被深度关注. 话题演化是网络舆情分析的重要内容之一，本文尝试从话题热度变化和内容变化两方面研究舆情动态. 本文选取天涯论坛民生讨论的主要版块——天涯杂谈的首发帖为舆情来源，分析比较一系列主题模型后，建立动态主题模型（DTM）. 通过挖掘随时间变化的动态话题链，从词语变化的微观角度分析热门事件下公众意见的变迁过程，还原事件的整个发展过程. 本文提出话题热度计算方法，通过计算 2012 全年天涯杂谈版块下所有新发帖的话题热度值变化及可视化分析，总结了 BBS 话题的三个规律.

关键词： 主题模型；DTM；话题演化；天涯论坛
中图分类号： C939 **文献标识码：** A **文章编号：** 1007－9807(2014)11－0109－13

0 引 言

互联网已成为人们沟通交流、分享信息、表达情感的重要场所，许多热门社会事件会在网络上形成强大的舆论动向，随时间不断演化发展. 比如“郭美美”事件，最早是 2011 年 6 月因新浪微博用户郭美美实名认证为“中国红十字会商业总经理”并在微博中炫富而引发的社会争议，后来随着网民的不断参与，矛头很快指向中国红十字会，最终引起一场慈善信任危机. 类似例子还有“毒胶囊事件”引发对政府监管的信任危机等. 由此可见，网络舆论会对现实中人们的生活带来深远的影响，同时也对政府的社会管理工作提出了更高的要求. 如何准确地把握网络舆论，并在适当的时间采取监管措施，是政府在进行社会管理时面临的实际问题.

在众多网络媒体中，论坛是一个重要的网络媒介，是反映网络舆情的一面“镜子”，正日益成为网络舆情的重要数据来源. 网络论坛具有时效性、价值体系多元化、平等互动性等特点[1]. 它比博客具备更强的可交互性，比微博更利于网民分析事件的发展演变过程. 虽然最近两年微博开始崭露头角，每天更新的数量要远远高于论坛，但对于突发热点事件而言，微博的信源一般并不明确，主要起到迅速提供消息的作用，而论坛在扩大突发舆情影响力方面的作用不可取代[2].

中国大陆大约 130 万个论坛，在其中一些颇有影响力的论坛上，网民就社会热点发表的言论对政府工作与决策显示出强大的影响. 在前期对天涯杂谈版块（天涯论坛的子版块）数据进行每日采集[3]的基础上，本文进一步聚焦每日话题变迁的研究. 这些帖子涵盖的话题不仅包括日常生活，更多的是对社会突发事件、社会不公、社会道德、反腐败等关乎民生的讨论，是研究社会舆情较好的数据源.

① 收稿日期：2014－05－19；修订日期：2014－09－16.
基金项目：国家重点基础研究发展计划资助项目（2010CB731405）；国家自然科学基金资助项目（71171187；71371107）.
作者简介：曹丽娜（1985—），女，河北邢台人，博士生. Email：caolina@amss.ac.cn
本文入选第十二届全国青年管理科学与系统科学学术会议优秀论文.

话题演化是网络舆情分析的重要内容之一，本文将从两个方面考察话题演化的表现，一是话题热度随时间的波动，话题热度(强度)是指对话题/事件本身的关注强度、讨论的激烈程度；二是话题内容本身的演化，用来体现话题视点的变迁.

本文尝试运用动态主题模型(dynamic topic models，DTM)对天涯杂谈版块中 2012 年全年所有首发帖建模，模型可从话题热度变化及内容变迁两方面共同解释舆情变化. 该研究为网络舆情分析提供新的有效途径，扩展模型应用范围. 对于管理决策者来说，文章更多地是从宏观视角呈现动态网络舆情，有助于把握舆情整体态势. 文章结构如下：首先介绍话题演化分析的现状，之后阐述数据的清理过程，论述对模型结果的分析处理过程，最后给出结论.

1 话题演化研究

TDT(topic detection and tracking)是最早用来研究话题演化的方法[4]. TDT 源于 1996 年美国国防高级研究计划委员会(DARPA)提出需要一种能自动确定新闻报道流中话题结构的技术. 随后，DARPA、卡内基·梅隆大学、Dragon 系统公司以及马萨诸塞大学的研究人员开启了这项领域的研究.

该研究对新闻报道进行处理(6 个任务)：报道切分、话题关联识别、新事件发现、话题追踪、话题发现、分层话题发现[5]. 从某些角度上说，TDT 各任务其实就是一种检索、过滤、分类的问题，与信息检索、信息过滤、文本分类等有一定的关联.

话题演化分析面临的挑战主要来自于文本的自然语言处理方面. 网络论坛帖子文本是由很多异构的文本组成，具有样本规模巨大、特征空间多维等特点，这对文本挖掘技术中最基本的特征空间维度表达以及相应的计算带来很大挑战，已成为当前制约网络文本挖掘研究的关键瓶颈问题之一，寻求新的技术成为必然. 解决高维度问题一般需要对文本进行降维，把高维的词空间映射到低维的语义空间. 近年来得到快速发展的基于概率图模型的主题模型[6]在处理文本数据时表现出较好效果.

1.1 主题模型

主题模型作为一种新的统计方法，它通过分析非结构化文本中的词语以发现蕴藏于其中的主题，然后利用获得的主题进行信息检索、分类、聚类、摘要提取以及信息间相似性、相关性判断等一系列应用[7]. 近年来，主题模型已逐渐成为文本挖掘、信息检索等领域的一个新的研究方向.

普林斯顿大学 Blei 教授等最早提出的 LDA(latent dirichlet allocation)模型是主题建模中最基本的模型[7]，之后又提出一系列基于 LDA 模型的改进模型. 其中一些模型弱化了 LDA 的统计假设，如弱化了词袋模型假设[8]、假设主题随时间变化[9]、假设主题相关等[10]. 另外还可以结合文档以外的作者、题目、链接等信息进行分析[11]，这些扩展的主题模型除了分析文本数据(如微博热门博主挖掘[12]、情绪挖掘[13]等)，针对其他数据类型(如社会网络、图像、源代码、生物信息等)也得到广泛应用.

1.2 基于主题模型的话题演化研究现状

相较于静态的文档集而言，时间序列文本中每篇文档所附带的时间信息对理解文本内容有非常大的帮助. 动态的主题模型假设主题随时间变化，可从时间序列文本中辨识和追踪动态的主题.

目前基于主题模型进行话题演化的研究主要是根据时间引入方式的不同分为以下三种：

1) 后离散方式

大多基于 LDA(latent dirichlet allocation)模型[7]，即先忽略时间，在整个文本集上运用 LDA 模型获取所有的话题，再按照文本的时间信息将文档离散到相应的时间点. 对于某个话题，可以依次考察它在每个时间点下的话题强度，从而在整个时间轴上显示出随时间推移话题的上升或下降. 该模型的优点是简单易行，但由于模型假设文档顺序是可交换的，没有将时间信息与模型有效结合，因此并未充分利用时间信息，导致同样建模条件下与其他动态主题模型相比会出现困惑度值很高的情况[14].

2) 引入时间变量方式

随时间变化的主题模型 (topic over time, TOT) 采用 Beta 分布对主题强度在给定时间范围内的变化进行建模,将文本、词项和时间三者联合起来作为观测数据[15].

连续时间动态主题模型 (continuous time dynamic topic model, cDTM) 使用布朗运动 (Brownian motion) 来模拟表示话题分布的参数在时间上的演化[16].

这两种方法的优点是模型的时间信息是连续的,不会出现时间粒度选取的问题. 不足的是 TOT 模型假设任意时刻的话题分布相互独立,而且忽略了话题内容的变化,仅仅展示话题强度的变化; cDTM 模型对数据本身有一定的要求,这是由于该模型在变分后验推断中充分运用了文本稀疏的特性,因此相对较"偏爱"变化大的数据,影响了模型的泛化能力.

3) 先离散方式

动态主题模型(dynamic topic models, DTM)[9] 先将时间离散化切片,各时间片下的主题概率分布和词语概率分布均依赖于前一个时间片的状态. 模型充分利用了话题之间的连续性,能同时分析话题强度和内容的变化,但是存在粒度选择的问题.

在线 LDA 模型(Online LDA, OLDA)[17] 根据时间信息将文本集划分为一组时间窗口,应用 LDA 模型发现每个时间窗口文本集的话题,并且采用话题历史分布作为当前时间窗口话题发现的先验知识,研究话题内容和强度的演化. 该模型基于在线随机优化技术,能够增量处理文本数据即流数据[18],但存在话题关联、话题探测的问题.

2　BBS 话题及其演化研究

目前,国外对于话题演化的分析多集中于对新闻[16]、科学文献[17]、电子邮件[15]、电影评论[19]等文本的挖掘与分析,很少有针对网络论坛/BBS 的帖子文本话题及其演化研究.

已有国内学者针对网络论坛的相关研究,如: 分析论坛发帖的点击量、回复量等行为寻找热点事件[20]; 挖掘 BBS 某个热点事件讨论中的意见领袖[21]; 采用 TDT 方法[22] 或类似方法[23] 挖掘 BBS 热点话题、追踪话题演化等. 这些研究都区别于本文基于主题模型对话题演化的分析.

胡艳丽等基于 1.2 节的 OLDA 模型研究了天涯社区经济论坛帖子的话题演化[24]. 石大文等采用 1.2 节中的方法 1 研究了 BBS 中的帖子[25],其中采用的是手工建立的语料集,只选取 11 个热门事件的 350 篇帖子展开分析. 它们与本文抽取一年天涯杂谈版块数据样本进行的分析有显著差异.

2.1　模型选择

天涯杂谈版块中的帖子相比新闻文本、科学文献而言非常不规范,具体表现为: 帖子内容的表达形式多样化,体裁形式多样化,文本长度差异性大,每日更新文本量大. 主题模型是对内容的分析,故对内容依赖性较强. 天涯杂谈的这些文本特点对主题模型的适应性、泛化能力提出了要求. 本文在模型选择时进行了以下分析和尝试.

本文曾尝试 1.2 节方法 1 中的 LDA 建模,并与 DTM 模型建模比较,发现同样数据样本下两个模型的结果比较相似,但 LDA 模型呈现出的话题热度更加平滑,话题分布比较均衡,一些热门事件并没有被很好地突出出来[26]. 1.2 节方法 2 讨论中已指出,cDTM 模型较适合于分析时间戳较密集、稀疏值(sparsity) 较高的文本[16]. 通过对以天为单位时间点的天涯杂谈帖子的稀疏性计算,发现该文本集稀疏值很低,更适合于 DTM 算法的建模. 另外因 TOT 模型并不支持对话题内容的演化分析,故放弃了对该模型的尝试.

综上,本文研究最终选择使用 DTM 模型建立动态主题链,根据文本—主题层级的时间戳标示,对主题分布做综合处理从而刻画话题总的热度及其热度变化. 同时探测话题—词语层级下词语的演变,发现视点的变迁.

2.2 DTM 模型

DTM 模型先将时间离散化切片，然后假定相邻时间片上整个文档的主题分布以及主题内容都是随时间演化的. 在 t 时刻，文档集的主题分布 α_t 以及主题下词语分布为 $\beta_{t,k}$ 均依赖于上一时刻的 α_{t-1} 和 $\beta_{t-1,k}$. 在 t 时刻，文档的生成过程如下[9]：

1) 抽取词语的分布 $\beta_t \mid \beta_{t-1} \sim N(\beta_{t-1}, \sigma^2 I)$.

2) 抽取主题分布 $\alpha_t \mid \alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$.

3) 针对每篇文档:

(1) 抽取主题分布 $\eta \sim N(\alpha_t, a^2 I)$.

(2) 针对每个词语:

i. 抽取主题 $Z \sim Mult(\pi(\eta))$

ii. 抽取词语 $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$

其中函数 π 是多项式到正态分布的映射. 即

$$f = \pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})},$$

$$\theta = \pi(\eta) = \frac{\exp(\eta)}{\sum_i \exp(\eta_i)}.$$

DTM 中如何根据观测值 $W_{t,d,n}$ 学习参数构成推断问题. 这里，潜在主题变量 $\theta_{d,t}$、$\beta_{t,k}$、$z_{t,d,n}$ 均是待估参数，文献[9]提出采用变分法估计参数.

3 数据及数据清理

天涯论坛(天涯社区)是中国的一个网络社区，截止2013年7月，注册用户达8 200万，在线用户常在80至100万左右. 在这里不仅诞生了一大批网络名人，也是传统媒体倚重的新闻源，许多事件一经过天涯论坛曝光就引起全社会的关注，是社会事件的催化剂和放大器.

3.1 天涯杂谈版块特点

在众多版块中，天涯杂谈帖子涵盖的话题范围较广，尤其对社会性话题的关注在天涯所有版块里首屈一指. 分析这类帖子数据能够一定程度上帮助了解热点事件和舆情走向. 虽然研究的并非网络全样本数据，但因天涯论坛的影响力和代表性，这些数据源已能较好地涵盖华语圈发生的各种事件.

3.2 数据清理

中科院系统所综合集成与知识科学研究小组从2010年10月份开始从包括“天涯杂谈”版块在内的若干个版块上获取每日首发帖和更新帖[3]. “天涯杂谈”版块的每日首发帖量约在1 500条左右，更新帖(即有回复的帖子)则达到3 500条左右. 截止2014年5月份，总的首发帖数据量已经抓取达到了160多万条.

本文抽取2012年全年的每日新发帖数据并作如下清理: 1) 去除只有链接但已被删除的帖子; 2) 用ICTCLAS② 对首发帖进行切词处理，同时只保留名词、动名词，生成语料库; 3) 去掉词频小于50次并且在所有文档中出现次数小于10次的词语，生成词典. 表1列出了清理后每个月帖子的个数、语料库及词典中的词数.

表1 数据集概况

Table 1 Summary of the data

数据统计 / 时间跨度	帖子数量 (documents)	语料库数量(万) (corpus)	词典数量 (tokens)
2012年1月	12 032	147	3 973
2012年2月	20 124	270	6 091
2012年3月	37 549	502	9 516
2012年4月	32 939	398	8 089
2012年5月	33 471	407	8 105
2012年6月	24 371	284	6 097
2012年7月	30 657	344	7 175
2012年8月	40 231	428	8 299
2012年9月	37 418	392	7 623
2012年10月	39 158	433	8 579
2012年11月	42 100	443	8 459
2012年12月	40 527	464	603

② 中文切词技术，网址: http://www.ictclas.org/.

4　天涯论坛上话题的演化分析

DTM 模型训练中时间粒度和时间跨度的选择对模型的结果至关重要. DTM 一个重要的假设就是每个时间点下的主题分布都依赖于上一时刻的主题分布,具备一定的连续性,且是平滑演化的,而论坛上讨论的话题有一些是更新相对较快的. 为此尝试了一系列的模型实验: 时间跨度分别选择 1 个月、3 个月、4 个月去训练模型,同时用 1 天、3 天和 7 天为时间戳. 结果发现,与科学文献不同的是,因 BBS 上发帖围绕的是日常生活,若时间跨度较长,会"忽略"掉一些突然出现又很快消失的热点事件,只"抓住"一些持续性比较强的话题,换而言之,DTM 模型"不喜欢"变化太大的内容. 另外,若粒度过小、时间戳过多,模型训练会非常消耗时间和内存,严重影响效率,粒度过大则与 DTM 的一个重要假设(DTM 假设某时间片下的文档是可交换的) 发生冲突. 故本研究最终选择以月为时间跨度,分别训练 12 个模型,每个模型的时间戳为 1 天. 每个模型的主题个数设为 60 个.

4.1　模型参数估计

本文采用 Blei 发布的 DTM 源码③进行后验估计,运行的硬件环境为英特尔 Core i7-2 600 3.40GHz 四核处理器,8GB 内存,32 位 Linux 操作系统. 以 2012 年 6 月份的数据量大小为例,训练一个 60 个主题的模型大约耗时 6.5 个小时.

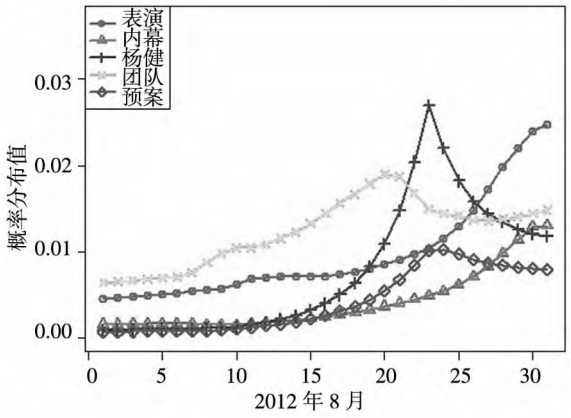
通过估计,得到了参数 $\beta_{t,k}$ 的分布(词语在 t 时刻下主题 k 中的对数分布) 、参数 $\theta_{d,t}$ 的分布(标签为时刻 t 的文本 d 的主题混合分布) 以及模型的对数似然估计的下界值(由 Jensen 不等式计算) 等参数值.

采用 R 语言读取上述参数结果并解析从而进行下一步分析.

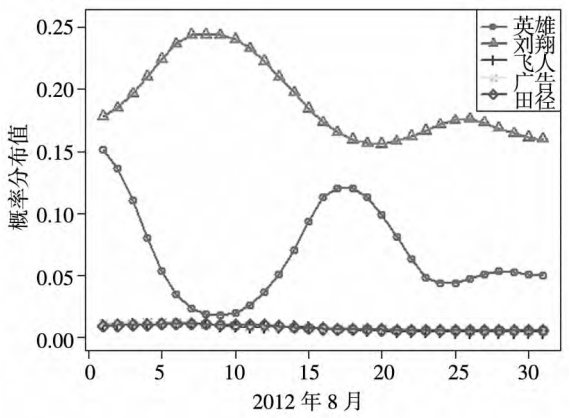
4.2　词语变化趋势

人们在不同时刻会谈论不同的话题,即便是针对同一个话题,所用词语也会有相应的变化. 根据后验参数 $\beta_{t,k}$ 的分布,在文档集这个层级,每个主题都由一系列词语的分布构成,同一主题在不同时间点下词语的分布也不同. 词语分布随时间的变化反映了每个话题中心的变迁.

本文从 2012 年 8 月的话题中选取"刘翔比赛"话题为这一分析做详细说明. 图 1 和表 2 从两个方面来展示在这个话题下词语随时间的变化: 图 1(a) 和图 1(b) 显示的是该话题下热门词语和冷门词语的变化. 其中 X 轴为时间轴,Y 轴为词语的概率分布值. 热门词语抽取的是月末和月初的概率正差值排名靠前的词语,即 Top ($\beta_{31,k} - \beta_{1,k}$) ,是热度值上升最大的几个词语. 相应地,冷门词语抽取的是在月初和月末的概率正差值排名靠前的词语,即 Top($\beta_{1,k} - \beta_{31,k}$) ,是热度值下降最多的词语. 表 2 抽取了若干个时间戳,给出每个时间戳下概率值排名前十的词语.



(a) 话题"刘翔比赛" 下的热门词语



(b) 话题"刘翔比赛" 下的冷门词语

图 1　"刘翔比赛" 话题下词语的变化趋势

Fig. 1 The trends of words in the topic "LIU Xiang"

③　http://code.google.com/p/princeton-statistical-learning/downloads/detail?name = dtm_release-0.8.tgz.

表 2 “刘翔比赛”话题下的前几个词语

Table 2 Several top words on the topic “LIU Xiang”

时间戳	话题“刘翔比赛”下的前十个词语
8 月 1 日	刘翔,英雄,比赛,广告,运动员,飞人,田径,伤病,成绩,冠军
8 月 4 日	刘翔,英雄,比赛,运动员,广告,飞人,田径,伤病,成绩,终点
8 月 7 日	刘翔,比赛,英雄,运动员,广告,飞人,田径,伤病,赛场,终点
8 月 10 日	刘翔,比赛,英雄,运动员,跟腱,赛场,团队,田径,广告,时间
8 月 13 日	刘翔,英雄,比赛,运动员,团队,终点,时间,赛场,田径,冠军
8 月 16 日	刘翔,英雄,比赛,团队,运动员,终点,时间,体育,国人,广告
8 月 19 日	刘翔,英雄,比赛,团队,运动员,终点,体育,杨健,表演,时间
8 月 22 日	刘翔,英雄,比赛,杨健,团队,运动员,终点,表演,体育,伤情
8 月 25 日	刘翔,英雄,比赛,杨健,团队,运动员,表演,终点,伤情,状态
8 月 28 日	刘翔,英雄,比赛,表演,运动员,团队,杨健,终点,内幕,体育
8 月 31 日	刘翔,英雄,表演,比赛,运动员,团队,内幕,杨健,终点,体育

针对“刘翔比赛”话题,可从两个重要的时间点(8 月 7 日比赛当天和 8 月 23 日新闻发布会的召开) 来分析话题的内容变迁过程. 赛前,人们对刘翔比赛寄予了厚望,因此“英雄”、“飞人”、“冠军”这样的词汇较热; 比赛当天,随着刘翔摔倒,词语“英雄”跌至最低点,赛后公众对刘翔的宽容使得该词慢慢回升. 8 月 23 日,央视在北京召开伦敦奥运报道研讨会,透露杨健解说的内幕,无意间让人们发现对刘翔的摔倒是有预案的,至此,词语“英雄”再次回落至低点. 与此同时,因公众对刘翔事件的质疑开始增加,“表演”、“内幕”等词汇开始变热.

4.3 主题变化趋势

DTM 模型假设某时刻下,文本—主题的分布条件依赖于上一时刻的分布. 根据后验参数 $\theta_{d,t}$ 的分布,可以从文档层级得到每个帖子中包含的主题分布. 为从整体把握各个话题热度的变化,需要采用一定方法来计算话题强度,定量描述话题的演化趋势.

话题强度的计算可参考斯坦福大学 Griffths 等[27] 和 Hall 等[28] 在后离散方式(基于 LDA 模型) 中提出的两种方法,前者采用取平均热度的方式,后者通过计算话题在文档中出现的次数得到.

对于 DTM 模型来说,$\theta_{d,t}$ 大多不为 0,按照 Hall 等人统计话题次数的方法没有意义. 这里参考 Griffths 等人的平均热度计算方法,将该思想应用至 DTM 模型的结果分析中. 具体方法如下:

记 $\theta_{d,t}$ 为帖子 d 在时刻 t 下的主题分布,平均的 $\theta_{k,t}$(记为$\bar{\theta}_{k,t}$) 表示主题 k 在时间 t 下的热度/强度,则$\bar{\theta}_{k,t}$ 的波动反映了公众对该话题关注度的变化. $\bar{\theta}_{k,t}$ 的具体计算过程为:

步骤 1 根据时间 $t = 1,\cdots,T$,重复如下步骤:

步骤 1.1 统计时间点 t 下的帖子个数 M_t;

步骤 1.2 在 t 时刻,第 k 个主题的热度值取为: $\bar{\theta}_{k,t} = \frac{1}{M_t}\sum_{d}\theta_{d,t}$;

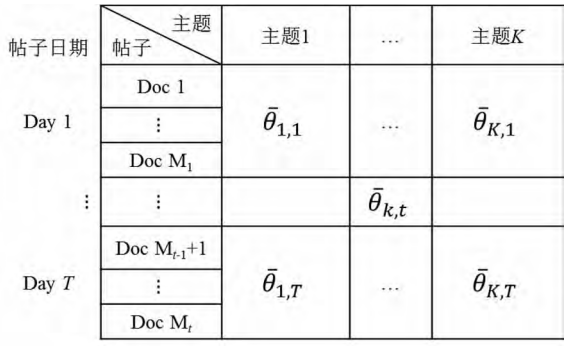


图 2 计算原理示意图

Fig. 2 Diagram of the calculation

图 2 是该计算原理的示意图. 通过计算得到 2012 年各月首发帖话题热度值 $\bar{\theta}_{k,t}$,如图 3 所示,包括 12 幅子图,每个子图中的横轴为时间,纵轴为话题. 这里只按总热度值(即 $\sum_{t=1}^{T}\bar{\theta}_{k,t}$) 排名由高到低显示其中 30 个总热度较高的话题(去掉了噪声话题),话题标签为人工总结给定.

图 3 子图中每个小单元格代表时间 t 下主题 k 的话题强度 $\bar{\theta}_{k,t}$,颜色表示话题强度值 $\bar{\theta}_{k,t}$ 所在的区间,颜色越深则代表热度值越高. 颜色区间所在的值域范围为: $[\min(\bar{\theta}_{k,t}),\max(\bar{\theta}_{k,t})]$,其中 $t \in (1,T)$,$k \in (1,30)$. 每个子图的图例给出了不同颜色所对应的具体热度区间值(各月模型不同,热度值区间也不同).

文后附录选取了一些比较有代表性的话题. 由于同一话题在不同时间点下词语的相似性较高，因此只截取了单个时间戳下的词语进行展示. 这些可以帮助理解天涯杂谈的话题内容.

图 3 直观呈现了 2012 年各月首发帖话题热度的变化，进而探寻到一些变化规律. 综合来看，天涯杂谈网络舆情具备的特点为:

1) 一些常规话题(每月都有，波动不大) 相对持续保持一定的热度水平，这样的话题包括:

(1) 日常生活类，如人生、心情、家庭、婚姻、爱情、朋友、大学生、教育、工作等话题;

(2) 对社会类问题的讨论，如社会道德、社会公知、电话诈骗、消费者投诉等话题;

(3) 经济文化类，如中医、文化、经济发展、历史等话题;

(4) 政府执政相关类，如社会改革、法律法规、反腐、政府与百姓、干部与群众等话题;

(5) 弱势群体维权相关，如医患纠纷、土地拆迁、村民选举、派出所民警执法、法院判决案件等话题;

(6) 人口资源与环境类，包括人口政策、环境污染问题.

2) 一些与特定时间或节日相关话题的热度会随日期来临而逐渐变热，此类话题热度一定程度上是可预测的，包括:

(1) 对圣诞节、雷锋日、春节、毛泽东诞辰等相关话题的讨论. 如 9 月 10 日教师节，“老师与学生”的话题热度会有小幅增长;

(2) 由节日本身衍生而出的话题，如 1 月份因春节来临，关于春运、订票、春晚等话题在节日前后成为人们关注的中心;

(3) 特定日期的大型活动相关话题，如伦敦奥运会及其衍生话题(刘翔比赛、羽毛球比赛等) .

3) 突发事件、热点事件相关话题，包括公众人物言论或司法案件导致的热点事件. 尤其是当威胁到人们安全或国家利益的突发事件发生时，人们会第一时间发帖申诉，形成热点话题. 这类话题的热度变化表现为突增，且较热，但持续性不是太强.

(1) 国家层面突发事件，如中日冲突，中菲争端，民族问题等话题;

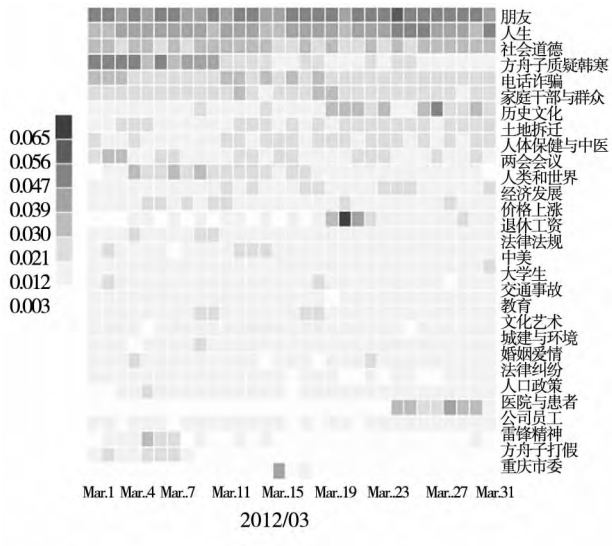
(2) 公众人物言论/行为类话题，如方韩大战、孔庆东事件、张绍刚事件等;

(3) 热门司法案件，如陶汝坤案、药家鑫案、周克华案等;

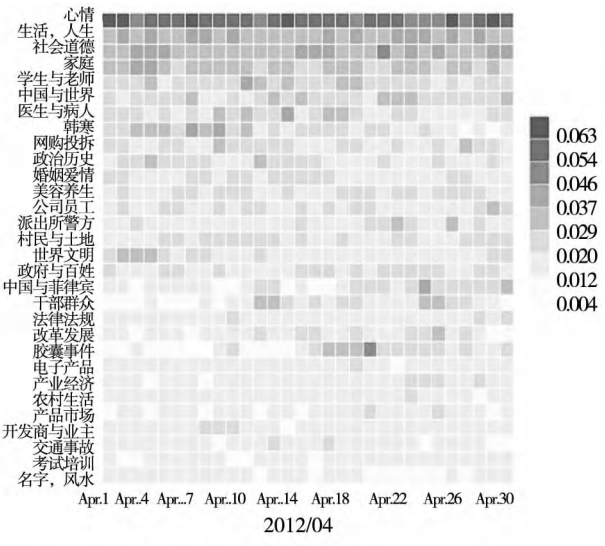
(4) 与群众利益/安全相关事件，如食品安全、毒胶囊事件、物价上涨、北京 7·21 城市暴雨等.

(1)

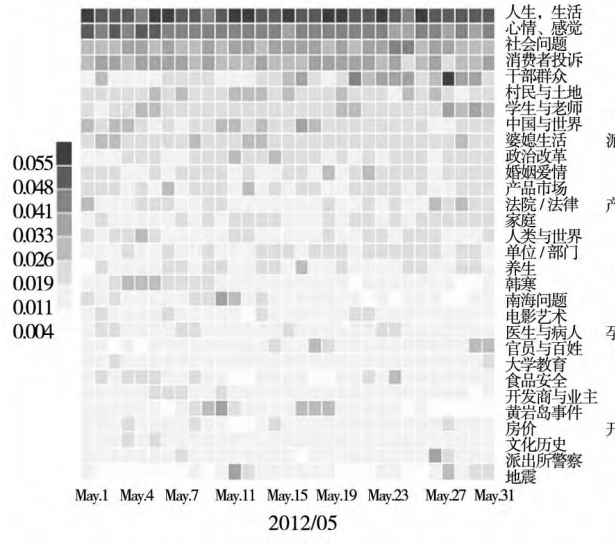
(2)


朋友
人生
社会道德
方舟子质疑韩寒
电话诈骗
家庭干部与群众
历史文化
土地拆迁
人体保健与中医
两会会议
人类和世界
经济发展
价格上涨
退休工资
法律法规
中美
大学生
交通事故
教育
文化艺术
城建与环境
婚姻爱情
法律纠纷
人口政策
医院与患者
公司员工
雷锋精神
方舟子打假
重庆市委
0.065
0.056
0.047
0.039
0.030
0.021
0.012
0.003
Mar.1 Mar.4 Mar..7 Mar.11 Mar..15 Mar..19 Mar..23 Mar..27 Mar.31
2012/03

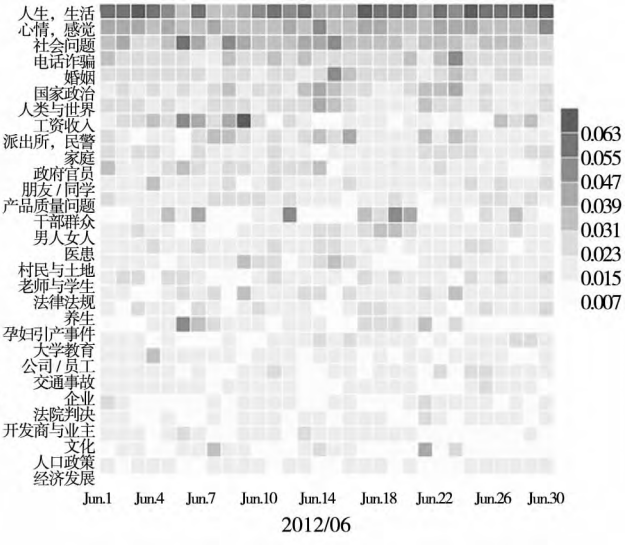
(3)


心情
生活，人生
社会道德
家庭
学生与老师
中国与世界
医生与病人
韩寒
网购投拆
政治历史
婚姻爱情
美容养生
公司员工
派出所警方
村民与土地
世界文明
政府与百姓
中国与菲律宾
干部群众
法律法规
改革发展
胶囊事件
电子产品
产业经济
农村生活
产品市场
开发商与业主
交通事故
考试培训
名字，风水
0.063
0.054
0.046
0.037
0.029
0.020
0.012
0.004
Apr.1 Apr.4 Apr..7 Apr..10 Apr..14 Apr.18 Apr.22 Apr.26 Apr.30
2012/04

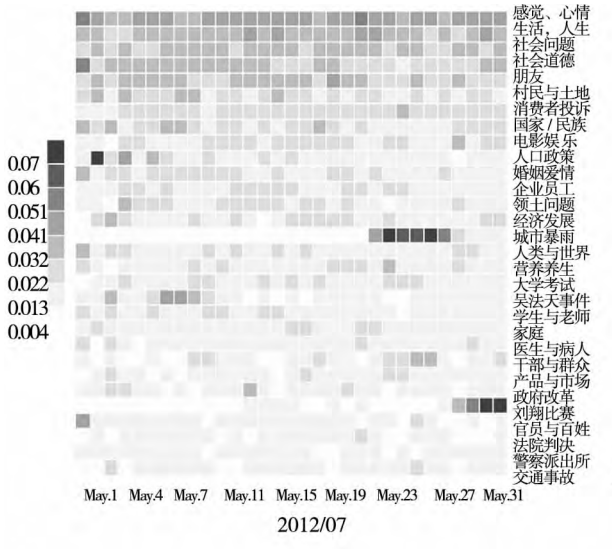
(4)


人生，生活
心情、感觉
社会问题
消费者投诉
干部群众
村民与土地
学生与老师
中国与世界
婆媳生活
政治改革
婚姻爱情
产品市场
法院/法律
家庭
人类与世界
单位/部门
养生
韩寒
南海问题
电影艺术
医生与病人
官员与百姓
大学教育
食品安全
开发商与业主
黄岩岛事件
房价
文化历史
派出所警察
地震
0.055
0.048
0.041
0.033
0.026
0.019
0.011
0.004
May.1 May.4 May.7 May.11 May.15 May.19 May.23 May.27 May.31
2012/05

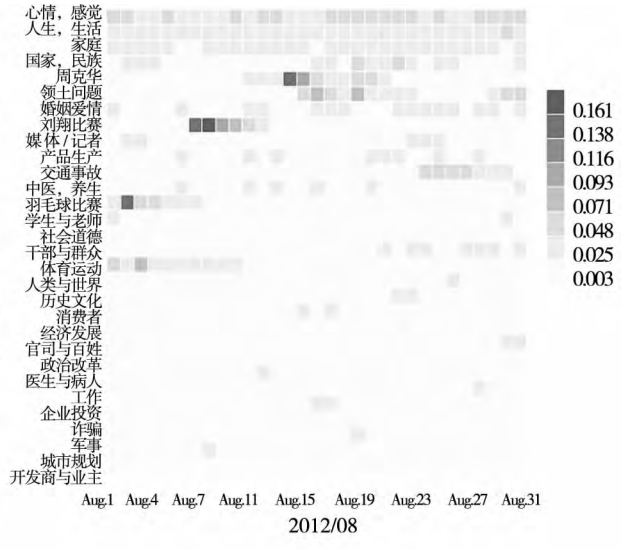
(5)


人生，生活
心情，感觉
社会问题
电话诈骗
婚姻
国家政治
人类与世界
工资收入
派出所，民警
家庭
政府官员
朋友/同学
产品质量问题
干部群众
男人女人
医患
村民与土地
老师与学生
法律法规
养生
孕妇引产事件
大学教育
公司/员工
交通事故
企业
法院判决
开发商与业主
文化
人口政策
经济发展
0.063
0.055
0.047
0.039
0.031
0.023
0.015
0.007
Jun.1 Jun.4 Jun.7 Jun.10 Jun.14 Jun.18 Jun.22 Jun.26 Jun.30
2012/06

(6)


感觉、心情
生活，人生
社会问题
社会道德
朋友
村民与土地
消费者投诉
国家/民族
电影娱乐
人口政策
婚姻爱情
企业员工
领土问题
经济发展
城市暴雨
人类与世界
营养养生
大学考试
吴法天事件
学生与老师
家庭
医生与病人
干部与群众
产品与市场
政府改革
刘翔比赛
官员与百姓
法院判决
警察派出所
交通事故
0.07
0.06
0.051
0.041
0.032
0.022
0.013
0.004
May.1 May.4 May.7 May.11 May.15 May.19 May.23 May.27 May.31
2012/07

(7)


心情，感觉
人生，生活
家庭
国家，民族
周克华
领土问题
婚姻爱情
刘翔比赛
媒体/记者
产品生产
交通事故
中医，养生
羽毛球比赛
学生与老师
社会道德
干部与群众
体育运动
人类与世界
历史文化
消费者
经济发展
官司与百姓
政治改革
医生与病人
工作
企业投资
诈骗
军事
城市规划
开发商与业主
0.161
0.138
0.116
0.093
0.071
0.048
0.025
0.003
Aug.1 Aug.4 Aug.7 Aug.11 Aug.15 Aug.19 Aug.23 Aug.27 Aug.31
2012/08

(8)

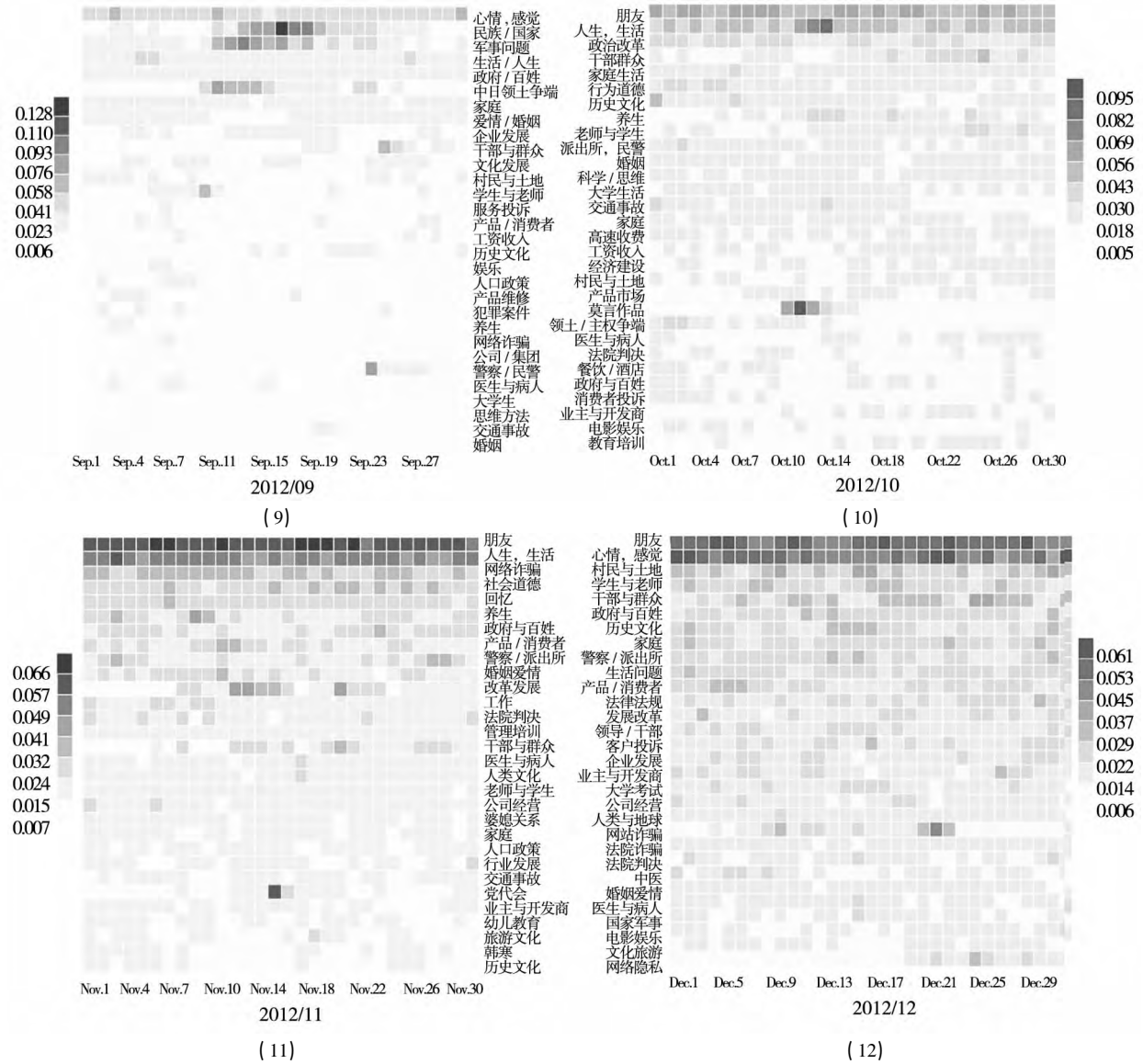


图 3　2012 年天涯杂谈版块每月话题热度变化

Fig. 3 The evolution of topics strength on Tianya Zatan board in each month of 2012

4.4　纵向话题热度变化

将一些持续时间比较长的话题从每个月中抽出来放到整个时间轴上，可以纵向观察到话题的波动. 这里相同话题的抽取是通过比较某个话题在不同时间戳下话题里词语的余弦夹角相似度找到其它月份中与其最相似的话题来作为该话题的延续.

图 4 和图 5 分别展示了话题“韩寒”和“政府与百姓”的热度变化，图中横线是平均值水平. 在图 4 中，方韩大战的发生使得对韩寒造假与否的争论成为热议，围绕“韩寒”的话题在 1 月底、2 月初达到讨论的峰值，并随着时间推移慢慢淡化.

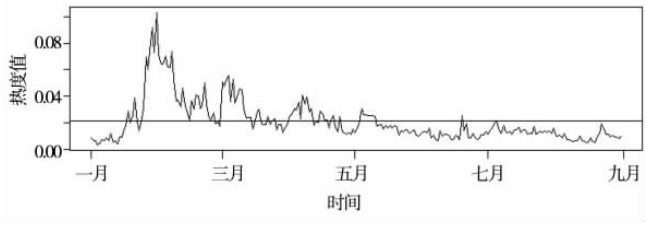


图 4　话题“韩寒”的热度变化

Fig. 4 The fluctuation of the strength of topic “HAN Han”

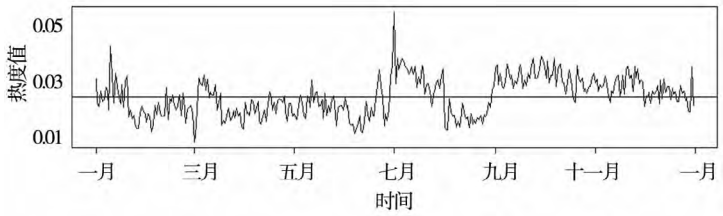


图 5 话题“百姓与政府”的热度变化

Fig. 5 The fluctuation of the strength of topic “government and common people”

图 5 中“政府与百姓”的话题则一直相对保持平稳，在春节期间降至较低水平，在 7 月份(7·21北京特大暴雨事件导致对政府城市建设的话题增多)高出平均值较多，而随着 8 月份伦敦奥运会的召开，民众注意力聚焦比赛，该话题热度开始降低.

5 结束语

互联网是反映社会舆论的主要载体之一. 有效把握网络舆情动态已成为国家权威机构的重要议题. 对政府来说，客观整体把握网络舆情，能够在掌握话题来龙去脉的同时了解一般民众的舆论诉求，为科学决策提供支持，从而正确引导舆论，维护国家稳定.

本文从中国大陆颇具影响力的天涯论坛切入，在分析论坛发帖特点后尝试用动态主题建模方法，定量化描述网络舆情温度变化，感知网络舆情规律特点. 实验结果表明，该方法能够有效检测话题演化，为网络舆情分析提供有效途径，贴合当前用自然科学方法定量化、模型化研究社会科学问题的发展趋势. 该模型也存在一些不足，比如因模型的复杂性导致模型的学习过程效率不够高.

BBS 这种开放式平台上，每条发帖形成一个即时研讨室，有的研讨室谈论热烈会持续相当长的时间，有的可能少有问津，它们都可视为 BBS 舆情研讨厅的研讨室. 综合集成与知识科学小组以往针对群体研讨的研究基本针对同一研讨室或者主题下群体研讨过程的详细分析，包括针对特定话题的群体意见挖掘所提出的基于不同机理的有效技术——CorMap 和 iView 分析话题变迁、探索群体思想结构[29]. 而 BBS 上的研讨，因研讨室的数量和规模在尺度上的放大，此时分析更关注整体话题的结构，故采用了主题模型来应对研讨尺度变化所带来的挑战.

未来工作会考虑将帖子内容分析与发帖行为分析结合，挖掘热门贴的话题内容与点击量、回复量之间的关系. 此外，针对天涯杂谈及其他天涯论坛社会及民生相关板块的话题分析也可作为一项常规工作，尝试每日、每周或每月的分析，以更好地应对社会管理创新的需求和挑战.

参 考 文 献:

[1] 尤薇佳，李 红，刘 鲁. 突发事件 Web 信息传播渠道信任比较研究[J]. 管理科学学报，2014，17(2)：19－33.
You Weijia，Li Hong，Liu Lu. Comparison of web channels for unconventional emergency events information dissemination [J]. Journal of Management Sciences in China，2014，17(2)：19－33. (in Chinese)

[2] 戴 玉. 李刚事件的早期传播流及舆论主题变化研究[D]. 济南：山东大学，2012.
Dai Yu. Research on the early transmission flow and changing of Li Gang case [D]. Jinan：Shandong University，2012. (in Chinese)

[3] 张泽代，唐锡晋. 面向天涯论坛的 Web 挖掘的初步研究[C]. 第十一届全国青年系统科学与管理科学学术会议暨第七届物流系统工程学术研讨会论文集，武汉：武汉理工大学出版社，2011：199－204.
Zhang Zedai，Tang Xijin. A Preliminary Study of Web mining for Tianya Forum [C]. Proceedings of the 11th Youth Conference of Systems Science and Management Science and 7th Conference of Logistic Systems Technology，Wuhan：Wuhan Uni-

versity of Science and Engineering Press, 2011: 199 – 204. (in Chinese)

[4] Allan J, Carbonell J G, Doddington G, et al. Topic detection and tracking pilot study final report [R]. Proceedings of the Broadcast News Transcription and Understanding Workshop (Sponsored by DARPA), February, 1998.

[5] Allan J. Introduction to Topic Detection and Tracking [M]. // In: J. Allan (eds.), Topic Detection and Tracking, Boston: Kluwer Academic Publisher, 2002, pp1 – 16.

[6] Blei D M. Probabilistic topic models [J]. Communications of the ACM, 2012, 55(4): 77 – 84.

[7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993 – 1022.

[8] Wallach H. Topic modeling: Beyond bag of words [C]. In: W. W. Cohen & A. Moore (eds.), Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, June 25 – 29, 2006: 977 – 984, ACM.

[9] Blei D M, Lafferty J D. Dynamic topic models [C]. In: W. W. Cohen & A. Moore (eds.), Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, June 25 – 29, 2006: 113 – 120, ACM.

[10] Blei D M, Lafferty J D. A correlated topic model of science [J]. Annals of Applied Statistics, 2007, 1: 17 – 35.

[11] 徐　戈, 王厚峰. 自然语言处理中主题模型的发展 [J]. 计算机学报, 2011, 34(8): 1423 – 1436.
Xu Ge, Wang Houfeng. The development of topic models in natural language processing [J]. Chinese Journal of Computers, 2011, 34(8): 1423 – 1436. (in Chinese)

[12] Weng J S, Lim E P, Jiang J, et al. Twitter rank: Finding topic sensitive influential Twitterers [C]. In: B. Davison, T. Suel, N. Craswell, et al. (eds.), Proceedings of the 3rd International Conference on Web Search and Web Data Mining, New York, February 4 – 6, 2010: 261 – 270, ACM.

[13] Mei Q, Ling X, Wondra M, et al. Topic sentiment mixture: Modeling facets and opinions in weblogs [C]. Proceedings of the 16th International Conference on World Wide Web, Alberta, 2007: 171 – 180, ACM.

[14] Iwata T, Yamada T, Sakurai Y, et al. Online multiscale dynamic topic models [C]. In: B. Rao, B. Krishnapuram, A. Tomkins, et al. (eds.), Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, July 25 – 28, 2010: 663 – 672, ACM.

[15] Wang X, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends [C]. In: T. E. Rad, L. H. Ungar, M. Craven, et al. (eds.), Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, August 20 – 23, 2006: 424 – 433, ACM.

[16] Wang C, Blei D, Heckerman D. Continuous time dynamic topic models [C]. In: D. A. McAllester & P. Myllym (eds.), Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, July 9 – 12, 2008: 579 – 586, AUAI Press.

[17] Alsumait L, Barbara D, Domeniconi C. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking [C]. Proceedings of the 8th IEEE International Conference on Data Mining, Pisa, December 15 – 19, 2008: 3 – 12, IEEE.

[18] Hoffman M D, Blei D M, Bach F R. Online learning for latent Dirichlet allocation [C]. In: J. D. Lafferty, C. K. I. Williams, J. S. Taylor, et al. (eds.), Proceedings of 24th Annual Conference on Neural Information Processing Systems, Vancouver, B. C., December 6 – 9, 2010: 856 – 864, Curran Associates, Inc.

[19] Meng C, Zhang M, Guo W Q. Evolution of movie topics over time [EB/OL]. http://cs229.stanford.edu/proj2012/MengZhangGuo-EvolutionofMovieTopicsOverTime.pdf (April – 1 – 2014).

[20] Chen X, Li J, Li S, et al. Hierarchical activeness state evaluation model for BBS network community [C]. In Proceedings of 7th International ICST Conference on Communications and Networking in China, Kunming, August 8 – 10, 2012: 206 – 211, IEEE.

[21] Zhou X, Yang J, Zhang J, et al. A BBS opinion leader mining algorithm based on topic model [J]. Journal of Computational Information Systems, 2014, 10(6): 2571 – 2578.

[22] Shi L, Sun B, Kong L, et al. Web forum sentiment analysis based on topics [C]. Proceedings of Ninth IEEE International Conference on Computer and Information Technology, Xiamen, October 11 – 14, 2009, 2: 148 – 153, IEEE.

[23] You L, Du Y, Ge J, et al. BBS based hot topic retrieval using back-propagation neural network [C]. In K. Y. Su, J. Tsujii, J. Lee, et al. (eds.), Proceedings of the 1st International Joint Conference on Natural Language Processing, Hainan Island, March 22 – 24, 2004, Revised Selected Papers, LNCS 3248: 139 – 148, Springer.

[24] 胡艳丽, 白 亮, 张维明. 网络舆情中一种基于 OLDA 的在线话题演化方法 [J]. 国防科技大学学报, 2012, 01: 150 – 154.
Hu Yanli, Bai Liang, Zhang Weiming. OLDA-based method for online topic evolution in network public opinion analysis [J]. Journal of National University of Defense Technology, 2012, 01: 150 – 154. (in Chinese)

[25] 石大文, 张 晖. 基于 LDA 模型的 BBS 话题演化 [J]. 工业控制计算机, 2012, 25(05): 82 – 84.
Shi Dawen, Zhang Hui. LDA model-based BBS topic evolution [J]. Industrial Control Computer, 2012, 25(05): 82 – 84. (in Chinese)

[26] Cao L N, Tang X J. Analysis of dynamic topics evolution based on Tianya club [C]. In: S. Y. Wang, Y. Nakamori & W. L. Jin (eds.), Proceedings of the 14th International Symposium on Knowledge and Systems Sciences, Ningbo, October 25 – 27, 2013: 146 – 154, JAIST Press.

[27] Griffths T, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(Suppl 1): 5228 – 5235.

[28] Hall D, Jurafsky D, Manning C D. Studying the history of ideas using topic models [C]. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Hawaii , October 25 – 27, 2008: 363 – 371, ACL.

[29] 唐锡晋. 两个定性综合集成支持技术 [J]. 系统工程理论与实践, 2010, 30(9): 1593 – 1606.
Tang Xijin. Two supporting technologies for qualitative meta-synthesis [J]. Systems Engineering: Theory & Practice, 2010, 30(9): 1593 – 1606. (in Chinese)

Trends of BBS topics based on dynamic topic model

CAO Li-na, TANG Xi-jin

Academy of Mathematics and Systems Science, Chinese Academy Sciences, Beijing 100190, China

Abstract: Along with the increasingly prominent issues aroused via Internet, online opinions are more and more concerned by researchers. This paper attempts to study the evolution of online opinions from two aspects, hot degree and BBS content, respectively. The original posts in Tiaya Zatan board, which is a primary board on social topics in Tianya Club, are collected as a source of online public opinion. By comparisons of a series of topic models, the dynamic topic model (DTM) is chosen. After mining the dynamic topics, the evolutions of public opinions toward hot events are analyzed under a micro perspective of changing words, so as to outline the process of those events. A method is proposed to compute the hot degree of the topic, and then the hot degree of all topics on the Tianya Zatan board in 2012 are computed and visualized, from which three rules of the topics are summarized under a macro perspective.

Key words: topic models; dynamic topic model; topics evolution; Tianya club

附录：

天涯杂谈话题内容示例

话题	老板与员工	产品与消费者	法院/法律	韩寒造假	交通事故	中菲争端	消费者投诉	感觉/心情	社会道德
词语	公司	企业	法院	方舟子	事故	台湾	网站	时候	社会
	老板	产品	法律	造假	交警	中国	服务	朋友	生活
	员工	市场	起诉	事实	动车	大陆	问题	感觉	人们
	工作	价格	证据	学术	车辆	南海	客户	事情	思想
	工资	成本	人民	逻辑	肇事	战争	投诉	东西	世界
	上班	生产	法庭	证据	交通	菲律宾	交易	时间	能力
	经理	品牌	法官	论文	家属	帕劳	卖家	心情	问题
	同事	食品	律师	文章	司机	渔民	密码	地方	时代
	加班	安利	事实	打假	死者	国家	消费者	同学	精神
	时间	标准	鉴定	质疑	责任	问题	用户	眼睛	人性
	小时	消费者	案件	团队	1 月	部队	买家	结果	内心
	辞职	地沟油	司法	问题	交通事故	台北	店铺	日子	事情
	工厂	行业	被告	韩寒	安全	军事	账号	声音	道德
	企业	超市	诉讼	抄袭	生命	越南	东西	关系	成功
	办公室	中国	原告	攻击	死亡	海军	评价	样子	价值

话题	婚姻爱情	家庭	改革发展	干部与群众	政府与百姓	拆迁	村民与土地	历史文化	中国与世界
词语	男人	父亲	发展	人民	百姓	政府	村民	中国	中国
	女人	母亲	改革	群众	政府	违法	农民	中国人	国家
	女性	女儿	管理	干部	人民	规定	土地	文化	美国
	老公	儿子	建设	同志	官员	行为	农村	国家	世界
	男性	妈妈	工作	党员	国家	拆迁	政府	世界	全球
	婚姻	爸爸	制度	重庆	代表	部门	百姓	美国	日本
	爱情	父母	社会	中央	社会	法律	村里	日本	外国
	夫妻	爷爷	服务	队伍	利益	人民	耕地	历史	法国
	老婆	奶奶	问题	领导	问题	房屋	国家	民族	国际
	感情	家庭	体制	问题	贪官	合法	集体	中华	英国
	女生	生活	方面	胡耀邦	民众	开发商	书记	发展	韩国
	妻子	妻子	基础	作风	领导	补偿	干部	政治	国人
	家庭	弟弟	基层	国家	群众	强拆	镇政府	文明	印度
	对方	姐姐	活动	政策	地方	国家	领导	思想	亚洲
	生活	孩子	任务	基层	权力	规划	情况	革命	德国

话题	学生与老师	医生与病人	生活/人生	人类与世界	政治改革	周克华事件	食品安全	城市暴雨	刘翔比赛
词语	学生	医生	生活	人类	国家	周克华	产品	暴雨	刘翔
	老师	医院	人生	思想	人民	警方	食品	城市	英雄
	学校	患者	时候	世界	中国	警察	生产	救灾	比赛
	教育	病人	梦想	教育	改革	女友	标准	灾难	运动员
	教师	医疗	感觉	动物	政治	死者	质量	下水道	表演
	家长	检查	生命	能力	体制	便衣	胶囊	大雨	田径
	校长	家属	时间	理论	社会	民警	问题	洪水	终点
	孩子	问题	内心	上帝	制度	悍匪	检测	排水	体育
	校车	结果	机会	地球	青年	案件	药品	受灾	商业
	小学	护士	工作	思维	社会主义	张贵英	消费者	灾害	时间
	领导	病情	事情	宗教	政府	照片	记者	救援	时候
	中学	病房	现实	问题	总理	克华	明胶	系统	团队
	班主任	主任	人们	信仰	历史	分析	销售	积水	结果
	问题	癌症	社会	人们	革命	时间	市场	特大	伤病
	中毒	人民	微笑	社会	利益	身份证	工业	生命	成绩